

# Randomization-Assisted Model-Based Survey Sampling

Phillip S. Kott

---

The Model-assisted paradigm presently dominates survey sampling. Under it, randomization-based theory is treated as the only true approach to inference. Models are helpful only when choosing between randomization-based methods. We propose an alternative theoretical paradigm. Model-based inference, which conditions on the realized sample, is the focus of this approach. Randomization-based methods, which focus on the set of hypothetical samples that could have been drawn, are employed solely to provide protection against model failure. Although the choices made under the randomization-assisted model-based paradigm are often little different from those recommended by Särndal et al. (1992), the motivation is clearer. Moreover, the approach proposed here for variance estimation leads to a logically coherent treatment of finite-population and small-sample adjustments when they are needed.

KEY WORDS: Asymptotic; calibration; estimation strategy; anticipated variance; simultaneous variance estimator.

---

Phillip S. Kott is Chief Research Statistician, Research and Development Division, National Agricultural Statistics Service, 3251 Old Lee Highway, Room 305, Fairfax, Virginia, 22030. This paper was prepared for the Fourth Biennial International Conference on Statistics, Probability and Related Areas, June 2002, DeKalb, Illinois.

## I. Introduction

Särndal, Swensson, and Wretman (1992) did not coin the term “model-assisted survey sampling,” but their impressive text-book has brought that approach to sample-survey inference into the forefront of modern theory and practice. The approach treats randomization-based (usually called “design-based”) inference as the real goal of survey sampling, but employs models to help choose between valid randomization-based alternatives. Typically, one chooses a randomization-consistent regression estimation strategy that is model unbiased and has the smallest model-expected randomization mean squared error.

To estimate the variance of the chosen strategy, Särndal et al. recommend the weighted residual variance estimator. Oddly, the real theoretical advantage of this variance estimator over the more traditional randomization-based variance estimator is that it better estimates the *model variance* of the regression estimator (SSW, 1989; Kott 1990a), while still estimating the randomization mean squared error adequately. The use of this variance estimator suggests a different approach to survey sampling inference: randomization-assisted model-based. In that approach, the subject of this discussion, one treats model-based inference as the goal of survey sampling, but employs randomization methods to protect against inevitable model failure. The choices for the estimation strategy and variance estimator do not change in the typical large-sample-much-larger-population environment, but the motivation behind the choices becomes clearer. Moreover, principled finite-population and small-sample

adjustments present themselves when necessary.

We will focus at first on a particular estimation strategy: the randomization consistent regression estimator (often called the “generalized regression estimator” or “GREG”) under Poisson sampling. This estimator is, as the name implies, randomization consistent. More to the point, it is model unbiased. Randomization unbiasedness is a fairly useless property since it tell us what happens when we average over all possible samples. In practice, we know which sample we have drawn, so averaging over samples we didn’t draw makes little sense. That is why the randomization-assisted model-based paradigm is principally concerned with the model unbiasedness of a parameter estimator rather than its randomization bias. By restricting attention to randomization-consistent estimation strategies, we are simply assuring ourselves that even when the model fails the estimator will likely not be too far from what it is estimating.

Section 2 describes the randomization-consistent-regression-estimator-under-Poisson-sampling strategy, while Section 3 analyzes its randomization and model-based properties, most of which are well known. We revisit the Isaki-Fuller (1981) notion of the anticipated variance of an estimation strategy, but reverse its meaning. Rather than choosing a strategy that has a small model-expected (anticipated) randomization (true) variance. We advocate choosing one with a small randomization-expected (anticipated before sampling) model (conditional on the realized sample) variance.

Section 4 addresses variance estimation. The emphasis in the *simultaneous variance estimator* is on estimating the model variance of the estimation strategy. The

literature suggests this emphasis can result in better coverage estimates. As protection against model failure however, the same estimator provides a nearly randomization-unbiased estimator of randomization mean squared error. Section 5 discusses how to modify the simultaneous variance estimator when  $n$  is not that large.

Section 6 addresses alternative sampling designs. Of particular interest, is the question of when the regression estimator is randomization consistent and what happens when cross terms are added to the simultaneous variance estimator. Section 7 provides some concluding remarks.

## 2. The Regression Consistent Estimator Under Poisson Sampling

Suppose we want to estimate a population ( $U$ ) total,  $T = \sum_U y_k$  based on a sample ( $S$ ) of  $y$ -values. If the probability that population unit  $k$  is in the sample is  $\pi_k$ , then the simple expansion of  $T$  is  $t = \sum_S y_k / \pi_k$ . Another useful way to render  $t$  is as  $t = \sum_U y_k I_k / \pi_k$ , where  $I_k$  is a random variable equal to 1 when  $k \in S$  and 0 otherwise. This means  $E(I_k) = \pi_k$ . Under randomization-based inference the  $y_k$  are fixed constants, while the  $I_k$  are random variables. It is easy to see that  $t$  is a randomization-unbiased estimator of  $T$ ; that is,  $E_p(t) = T$ , where the subscript  $p$  denotes the expectation with respect to the  $I_k$  (this is a convention; the  $p$  derives from “probability sampling”).

The randomization variance of  $t$  is

$$\text{Var}_p(t) = E_p[(t - T)^2] = \sum_U (y_k / \pi_k)(y_i / \pi_i)(\pi_{ki} - \pi_k \pi_i),$$

where  $\sum_U$  denotes  $\sum_{k \in U} \sum_{i \in U}$  in this context, and  $\pi_{ki} = E(I_k I_i)$  is the joint selection probability of units  $k$  and  $i$ . When  $k = i$ ,  $\pi_{ki} = \pi_k$ . The randomization variance of  $t$  depends on how exactly the sample is drawn, and in particular of the joint selection probabilities.

Under Poisson sampling, each unit  $k$  is sampled independently of every other unit in the population. Consequently,  $\pi_{ki} = \pi_k \pi_i$  when  $k \neq i$ . This simplifies the randomization variance of  $t$  immensely:

$$\text{Var}_p(t) = \sum_U (y_k / \pi_k)^2 (\pi_k - \pi_k^2) = \sum_U (y_k^2 / \pi_k) (1 - \pi_k),$$

and leads to the simple unbiased randomization variance estimator:

$$\text{var}_p(t) = \sum_S (y_k / \pi_k)^2 (1 - \pi_k).$$

We will be principally concerned here with the estimation strategy that combines Poisson sampling with this regression estimator:

$$t_R = t + (\sum_U \mathbf{x}_k - \sum_S \pi_k^{-1} \mathbf{x}_k) (\sum_S c_k \pi_k^{-1} \mathbf{x}_k' \mathbf{x}_k)^{-1} \sum_S c_k \pi_k^{-1} \mathbf{x}_k' y_k, \quad (1)$$

where  $\mathbf{x}_k = (x_{k1}, \dots, x_{kQ})$  is a row vector of values known for all  $S$ ,  $c_k$  is a constant,  $\sum_U \mathbf{x}_k$  is known, and  $\sum_S c_k \pi_k^{-1} \mathbf{x}_k' \mathbf{x}_k$  is invertible.

The regression estimator in equation (1) is a very slight variation of the general

regression estimator (GREG) in Särndal, Swensson, and Wretman. (1992). A good review of regression estimators in the survey sampling context is Brewer (1994). The GREG is poorly named because it does not include purely model-based regression estimators.

The regression estimator in equation (1) can be rewritten as  $t_R = \sum_S a_k y_k$ , where  $a_k$  is the regression weight of  $k$ :

$$a_k = \pi_k^{-1} + (\sum_U \mathbf{x}_i - \sum_S \pi_i^{-1} \mathbf{x}_i)(\sum_S c_i \pi_i^{-1} \mathbf{x}_i' \mathbf{x}_i)^{-1} c_k \pi_k^{-1} \mathbf{x}_k'. \quad (2)$$

It is well known (and easy to see) that the  $a_k$  satisfy the *calibration equation*:  $\sum_S a_k \mathbf{x}_k = \sum_U \mathbf{x}_k$ .

### 3. Properties of the Estimation Strategy

The regression estimator,  $t_R$ , under Poisson sampling has both desirable randomization-based and model-based properties under mild conditions, as we shall see.

#### 3.1. Randomization-based Properties

The randomization-based properties of  $t_R$  are asymptotic (we use the more accurate modifier “randomization” in place of the often-used “design” throughout the text). That is to say, they depend on the expected sample size,  $n^*$ , being large. A

sufficient condition for an estimation strategy (an estimator coupled with a sampling design) to be randomization consistent is that its relative mean squared error should approach 0 as  $n^*$  grows arbitrarily large.

Let  $N$  be the population size of  $U$ . We want to entertain the possibility that  $O(n^*)$  is less than  $O(N)$ . Consequently, we assume the following as  $N$  and  $n$  grow arbitrarily large and  $Q$  *remains fixed*:

$$0 < L_y \leq \sum_U y_k^\delta / N < B_y < \infty, \quad \delta = 1, \dots, 8; \quad (3.1)$$

$$0 < L_{xq} \leq \sum_U x_{kq}^\delta / N < B_{xq} < \infty, \quad q = 1, \dots, Q; \delta = 1, \dots, 8; \quad (3.2)$$

$$0 < L_c \leq \sum_U c_k^\delta / N < B_c < \infty, \quad \delta = 1, \dots, 8 \quad (3.3)$$

$$0 < L_\pi \leq \sum_U [(N/n^*)\pi_k]^\delta / N < B_\pi < \infty, \quad \delta = 1, \dots, 8. \quad (3.4)$$

The relative randomization mean squared error of the expansion estimator,  $t$ , under Poisson sampling is  $\sum_U (y_k^2/\pi_k)(1-\pi_k)/(\sum_U y_k)^2 < \sum_U (y_k^2/\pi_k)/(\sum_U y_k)^2$ . Equations (3.1), (3.4), and Schwarz's inequality tell us that the numerator of this last expression is  $O(N^2/n^*)$ , while its denominator is  $O(N^2)$ . Thus, the relative randomization mean squared of  $t$  under Poisson sampling is  $O(1/n^*)$ , and the estimation strategy is randomization consistent. Furthermore, since  $E_p[(t - T)^2]/T^2 = O(1/n^*)$ ,  $(t - T)/T = O_p(1/\sqrt{n^*})$ , and  $t - T = O_p(N/\sqrt{n^*})$ , which we will often render as  $NO_p(N1/\sqrt{n^*})$ . Formally, this means  $(t - T)/N = O_p(N1/\sqrt{n^*})$

The regression estimator,  $t_R$ , from equation (1) under Poisson sampling and the assumptions in equation (3) is equal to  $t + NO_p(1/\sqrt{n^*})$ . This is because, using similar argument as above (and Schwarz's inequality repeatedly), the components of

$\sum_U \mathbf{x}_k - \sum_S \pi_k^{-1} \mathbf{x}_k$  are  $NO_p(1/\downarrow n^*)$ , while the absolute values of the components of  $\sum_S c_k \pi_k^{-1} \mathbf{x}_k' \mathbf{x}_k$  and  $\sum_S c_k \pi_k^{-1} \mathbf{x}_k' y_k$  are  $NO_p(1)$ . Thus, like  $t$ ,  $t_R$  is randomization consistent. Furthermore,  $(t_R - T)/T = O_p(1/\downarrow n^*)$ , and the relative mean squared error of  $t_R$  is  $O(1/n^*)$ .

Assuming, as we will from now on, that  $N^{-1}(\sum_U c_k \mathbf{x}_k' \mathbf{x}_k)$  is invertible, let

$\mathbf{B} = (\sum_U c_k \mathbf{x}_k' \mathbf{x}_k)^{-1} \sum_U c_k \mathbf{x}_k' y_k$ , and  $e_k = y_k - \mathbf{x}_k \mathbf{B}$ , so that  $\sum_U c_i \mathbf{x}_i' e_i = 0$ . We can now express the error of  $t_R$  as

$$\begin{aligned}
 t_R - T &= \sum_S a_i y_i - \sum_U y_i \\
 &= \sum_S a_i (\mathbf{x}_i \mathbf{B} + e_i) - \sum_U (\mathbf{x}_i \mathbf{B} + e_i) \\
 &= \sum_S a_i e_i - \sum_U e_i \\
 &= \sum_S e_i / \pi_i + (\sum_U \mathbf{x}_k - \sum_S \pi_k^{-1} \mathbf{x}_k) (\sum_S c_k \pi_k^{-1} \mathbf{x}_k' \mathbf{x}_k)^{-1} \sum_S c_i \pi_i^{-1} \mathbf{x}_i' e_i - \sum_U e_i \\
 &= \sum_S e_i / \pi_i + (\sum_U \mathbf{x}_k - \sum_S \pi_k^{-1} \mathbf{x}_k) (\sum_S c_k \pi_k^{-1} \mathbf{x}_k' \mathbf{x}_k)^{-1} (\sum_U c_i \mathbf{x}_i' e_i + O_p(N/\downarrow n^*)) - \sum_U e_i \\
 &= \sum_S e_i / \pi_i + (\sum_U \mathbf{x}_k - \sum_S \pi_k^{-1} \mathbf{x}_k) (\sum_S c_k \pi_k^{-1} \mathbf{x}_k' \mathbf{x}_k)^{-1} O_p(N/\downarrow n^*) - \sum_U e_i \\
 &= \sum_S e_i / \pi_i - \sum_U e_i + NO_p(1/n^*) .
 \end{aligned}$$

This tells us that the randomization mean squared error of  $t_R$  under Poisson sampling is dominated by  $\text{Var}_P(\sum_S e_k / \pi_k) = \sum_U (e_k^2 / \pi_k) (1 - \pi_k)$ . This is identical to the variance of the expansion estimator under Poisson sampling except that  $e_k$  has replaced  $y_k$ .



### 3.2. Model-based Properties

Suppose the  $y_k$  were random variables that satisfied the following model:

$$y_k = \mathbf{x}_k \beta + \epsilon_k, \quad (4)$$

where  $\beta$  is an unknown column vector,  $E(\epsilon_k | \mathbf{x}_k, I_k) = E(\epsilon_k \epsilon_i | \mathbf{x}_k, \mathbf{x}_i, I_k, I_i) = 0$  for  $k \neq i$ , and  $E(\epsilon_k^2 | I_k) = \sigma_k^2 = f(\mathbf{x}_k, \mathbf{z}_k) < \infty$ , where  $\mathbf{z}_k$  is a vector of values associated with  $k$ . The  $\sigma_k^2$  need not be known. Moreover, there is no reason (yet) why  $I_k$  cannot be a function of the components of  $\mathbf{x}_k$  and  $\mathbf{z}_k$ .

It is easy to see that as long as the regression weights satisfy the calibration equation,  $\sum_S a_k \mathbf{x}_k = \sum_U \mathbf{x}_k$ ,  $t_R$  will be model unbiased in the sense that  $E_e(t_R - T) = 0$ . Moreover, its model variance (as an estimator of  $T$ ) is

$$\begin{aligned} E_e[(t_R - T)^2] &= E_e[(\sum_S a_i y_i - \sum_P y_i)^2] \\ &= E_e[(\sum_S a_i \epsilon_i - \sum_P \epsilon_i)^2] \\ &= \sum_S a_i^2 \sigma_i^2 - 2 \sum_S a_i \sigma_i^2 + \sum_U \sigma_i^2. \\ &= \sum_S a_i^2 \sigma_i^2 - \sum_S a_i \sigma_i^2 - (\sum_S a_i \sigma_i^2 - \sum_U \sigma_i^2). \end{aligned}$$

When  $\sigma_i^2$  has the form  $\mathbf{x}_i \mathbf{h}$ , for some not-necessarily-specified vector  $\mathbf{h}$ , then  $\sum_S a_i \sigma_i^2 = \sum_U \sigma_i^2$ , and the model variance of  $t_R$  collapses to  $\sum_S (a_i^2 - a_i) \sigma_i^2$ . Alternatively, if we add to our asymptotic assumptions,

$$0 < L_{\sigma} \leq \sum_U \sigma_k^{2r} / N < B_{\sigma} < \infty, \quad r = 1, \dots, 4, \quad (3.5)$$

then one can see that the model variance of  $t_R$  is  $O(N^2/n^*) = (N^2/n^*)O(1)$ , while  $\sum_S a_i \sigma_i^2 - \sum_U \sigma_i^2$  is  $O_p(N/\downarrow n^*) = (N^2/n^*)O(\downarrow n^*/N)$ . Although we are interested in model-based expectations, we plan to invoke a large-sample, randomization-based equality. Model-based theory, at least as viewed here, does not deny the applicability of the law of large numbers to probability sampling. It simply resists taking averages (expectations) across all possible samples.

Our last equality suggests the following asymptotic approximation for the model variance of  $t_R$ :

$$E_e[(t_R - T)^2] \approx \sum_S a_i^2 \sigma_i^2 - \sum_S a_i \sigma_i^2, \quad (5)$$

which drops a  $O_p(N/\downarrow n^*) = (N^2/n^*)O(\downarrow n^*/N)$  term.

What about likewise replacing  $a_i^2$  by  $\pi_i^{-2}$  (and  $a_i$  by  $\pi_i^{-1}$ ) in equation (5)? Such a substitution would effectively drop  $(N^2/n^*)O_p(1/\downarrow n^*)$  terms. To see why, observe that

$$\begin{aligned} \sum_S a_i^2 \sigma_i^2 &= \sum_S \pi_i^{-2} \sigma_i^2 + 2 (\sum_U \mathbf{x}_i - \sum_S \pi_i^{-1} \mathbf{x}_i) (\sum_S c_i \pi_i^{-1} \mathbf{x}_i' \mathbf{x}_i)^{-1} \sum_S c_i \pi_i^{-2} \mathbf{x}_i' \sigma_i^2 + \\ &\quad (\sum_U \mathbf{x}_i - \sum_S \pi_i^{-1} \mathbf{x}_i) (\sum_S c_i \pi_i^{-1} \mathbf{x}_i' \mathbf{x}_i)^{-1} \sum_S \sigma_i^2 c_i^2 \pi_i^{-2} \mathbf{x}_i' \mathbf{x}_i (\sum_S c_i \pi_i^{-1} \mathbf{x}_i' \mathbf{x}_i)^{-1} (\sum_U \mathbf{x}_i - \sum_S \pi_i^{-1} \mathbf{x}_i)' \\ &= \sum_S \pi_i^{-2} \sigma_i^2 + O_p(N^2/[n^*]^{3/2}) \\ &= \sum_S \pi_i^{-2} \sigma_i^2 + (N^2/n^*)O_p(1/\downarrow n^*). \end{aligned}$$

In subsequent analyses, such asymptotic arithmetic will often be left to the reader.

Suppose finite-population correction matters. At the extreme,  $N = O(n^*)$ , and  $(N^2/n^*)O_p(1/\downarrow n^*)$  is of the same asymptotic order as the  $(N^2/n^*)O(\downarrow n^*/N)$  term dropped in equation (5). An alternative assumption allows the finite population to be *relatively large* (Kott 1990a), but still potentially matter:  $N \geq O([n^*]^{3/2})$ . Under this regime, equation (5) appropriately drops a  $(N^2/n^*)O_p(1/n^*)$  term, but replacing  $a_i^2$  by  $\pi_i^{-2}$  would effectively drop a larger,  $(N^2/n^*)O_p(1/\downarrow n^*)$ , term.

### 3.3. *Anticipated Variance*

The model variance of  $t_R$  is a function of the realized sample and does not depend at all on the sampling design. As noted in the previous section, it is  $O_p(N^2/n^*)$  under the (extended) asymptotic assumptions of equation (3). In fact, if we are willing to drop  $(N^2/n^*)O_p(1/\downarrow n^*)$  terms, the model variance can be approximated by

$$E_\epsilon[(t_R - T)^2] \approx \sum_S (\sigma_i^2/\pi_i^2)(1 - \pi_i).$$

The randomization expectation of the model variance of  $t_R$  is then

$$E_p\{E_\epsilon[(t_R - T)^2]\} \approx \sum_U (\sigma_i^2/\pi_i)(1 - \pi_i). \quad (6)$$

This quantity can be called the “anticipated variance” of  $t_R$ ; that is, the model variance anticipated before random sampling. The term is due to Isaki and Fuller (1982),

although equation (6) goes back considerably further in the literature. They use it to mean  $E_{\epsilon}\{E_p[(t_R - T)^2]\}$ , what that model anticipates the randomization mean squared error to be. The expectation operators can be switched, and the two concepts of anticipated variance coincide, when  $\epsilon_k$  and  $\epsilon_k^2$  are uncorrelated with  $I_k$  given  $\mathbf{x}_k$  and  $\mathbf{z}_k$ , where  $\sigma_k^2 = f(\mathbf{x}_k, \mathbf{z}_k)$ , as we have assumed. This is weaker than the requirement that the  $\epsilon_i$  and  $I_i$  be independent, as stated in Isaki and Fuller. Maintaining the latter condition would rule out designs where  $\pi_k \propto \sigma_k$  for some hypothesized  $\sigma_k^2$ . This selection probability rule minimizes the asymptotic anticipated variance on the right hand side of equation (6) for a fixed expected sample size,  $n^* = \sum_U \pi_i$ . Brewer (1963) makes a similar point.

From equation (6), we can also see that the anticipated variance of the randomization-consistent regression estimator is (asymptotically) a function of the unit selection probabilities but not the joint selection probabilities. Every design with the same unit selection probabilities produces a regression estimator with the same anticipated variance. If minimizing anticipated variance is the goal, then *there is no penalty from using Poisson sampling*.

#### 4. Simultaneous Variance Estimation

It is a simple matter to estimate the (approximate) model variance of  $t_R$  expressed in equation (5):

$$v = \sum_S (a_i^2 - a_i) r_i^2, \quad (7)$$

where  $r_i = y_i - \mathbf{x}_i \mathbf{b}$ , and  $\mathbf{b} = (\sum_S c_k \pi_k^{-1} \mathbf{x}_k' \mathbf{x}_k)^{-1} \sum_S c_k \pi_k^{-1} \mathbf{x}_k' y_k$ . Now

$$r_i = e_i - \mathbf{x}_i (\mathbf{b} - \beta) = e_i - \mathbf{x}_i (\sum_S c_k \pi_k^{-1} \mathbf{x}_k' \mathbf{x}_k)^{-1} \sum_S c_k \pi_k^{-1} \mathbf{x}_k' e_k,$$

so

$$E(r_i^2) = \sigma_i^2 + 2\mathbf{x}_i (\sum_S c_k \pi_k^{-1} \mathbf{x}_k' \mathbf{x}_k)^{-1} c_i \pi_i^{-1} \mathbf{x}_i' \sigma_i^2 + \\ \mathbf{x}_i (\sum_S c_k \pi_k^{-1} \mathbf{x}_k' \mathbf{x}_k)^{-1} (\sum_S c_k^2 \sigma_k^2 \pi_k^{-2} \mathbf{x}_k' \mathbf{x}_k)^{-1} (\sum_S c_k \pi_k^{-1} \mathbf{x}_k' \mathbf{x}_k)^{-1} \mathbf{x}_i',$$

After a little work, we can conclude that  $v$  is asymptotically model unbiased when the population is relatively large,  $N \geq O_p([n^*]^{3/2})$ , so  $E_e[(t_R - T)^2] \approx \sum_S a_i^2 \sigma_i^2 - \sum_S a_i \sigma_i^2$ , and

$$E_e(v) = \sum_S (a_i^2 - a_i) \sigma_i^2 + (N^2/n^*) O_p(1/n^*). \quad (8)$$

Observe that the term we are ignoring in equation (8) are smaller than the  $(N^2/n^*) O_p(1/n^*)$  term we would have ignored had we replaced  $a_i$  with  $\pi_i^{-1}$ .

We can likewise show that  $v$  is an asymptotically unbiased estimator for the randomization mean squared error of  $t_R$  under Poisson sampling. In this context, however, we *are* willing to drop  $O_p(N^2/[n^*]^{3/2})$  terms. The equalities

$$\begin{aligned} r_i &= e_i - \mathbf{x}_i(\mathbf{b} - \mathbf{B}) = e_i - \mathbf{x}_i(\sum_S c_k \pi_k^{-1} \mathbf{x}_k' \mathbf{x}_k)^{-1} \sum_S c_k \pi_k^{-1} \mathbf{x}_k' e_k \\ &= e_i - \mathbf{x}_i \mathbf{O}_{(Q \times Q)p}(1/N) \mathbf{O}_{(Q \times 1)p}(N/\downarrow n^*), \\ &= e_i - O_p(1/\downarrow n^*) \end{aligned} \tag{9}$$

ultimately imply that  $v = \sum_S (a_i^2 - a_i) r_i^2 = \sum_S (\pi_i^{-2} - \pi_i^{-1}) e_i^2 + O_p(N^2/[n^*]^{3/2})$ . From which, we conclude

$$E_p(v) = \sum_U (\pi_i^{-1} - 1) e_i^2 + (N^2/n^*) O_p(1/\downarrow n^*). \tag{10}$$

We call  $v$  the *simultaneous variance estimator* because it simultaneously estimates the model variance and randomization mean squared error of  $t_R$ . The *relative* model bias of  $v$  (as an estimator of  $E_e[(t_R - T)^2] \approx \sum_S a_i^2 \sigma_i^2 - \sum_S a_i \sigma_i^2$ ) is  $O_p(1/n^*)$  when the population is relatively large; see equation (8). Its relative randomization bias (as an estimator of  $E_p[(t_R - T)^2] \approx \sum_U (\pi_i^{-1} - 1) e_i^2$ ) is  $O(1/\downarrow n^*)$ ; see equation (10). Empirical analyses like that in Wu and Deng (1983) have showed that this emphasis on the making the model bias small, which is the core of the randomization-assisted model-based paradigm, can lead to superior coverage estimates.

## 5. Adjusting for Small-sample Bias

It is tempting to scale  $v$  in equation (7) by  $n/(n - Q)$  to account for the fact that  $r_k^2$ , a squared residual from a  $Q$ -variate regression, is a slightly biased estimator for  $\sigma_k^2$ . Since the factor,  $n/(n - Q)$ , is asymptotically unity, the scaling does not affect the randomization-based properties of  $v$ .

A more principled approach than the above *ad-hoc* adjustment of  $v$  would be to replace the  $r_i^2$  with model unbiased estimators for the components of  $\sigma^2 = (\sigma_1^2, \dots, \sigma_n^2)$ , namely,  $\mathbf{r}_{(2)} = \mathbf{M}^{-1} (r_1^2, \dots, r_n^2)'$ , where the  $i, k$ th element of the  $n \times n$  matrix  $\mathbf{M}$  is  $m_{ik} = [\delta_{ik} - \mathbf{x}_i (\sum_S c_j \pi_j^{-1} \mathbf{x}_j' \mathbf{x}_j)^{-1} c_k \pi_k^{-1} \mathbf{x}_k]^2$ , and  $\delta_{ik} = 1$  when  $i = k$ , 0 otherwise. See Chew (1970). Calculating  $\mathbf{r}_{(2)}$  involves inverting an  $n \times n$  matrix. Kott and Brewer (2001) show how  $\mathbf{r}_{(2)}$  can be calculated by inverting a  $Q(Q+1)/2 \times Q(Q+1)/2$  matrix instead. Replacing the  $r_i^2$  by the components of  $\mathbf{r}_{(2)}$  does not affect the randomization consistency of  $t_R$  because  $\mathbf{M}$  is asymptotically the identity matrix.

A simpler alternative relies on assuming that  $s_k^2 \propto \sigma_k^2$  for some known  $\mathbf{s}^2 = (s_1^2, \dots, s_n^2)'$ . One replaces each  $r_i^2$  in  $v$  with the model-unbiased estimator:

$$r_{iA}^2 = r_i^2 s_i^2 / E(r_i^2 \mid \sigma^2 = \mathbf{s}^2) = r_i^2 s_i^2 / \sum_{k \in S} m_{ik} s_k^2. \quad (11)$$

This can produce an exactly model-unbiased variance estimator, call it  $v_A$ , when the assumption  $s_k^2 \propto \sigma_k^2$  – called “the working model” – is correct, but not generally. Still, it

has the same desirable asymptotic model and randomization-based properties as  $v$  when equation (3.5) is assumed to apply to the  $s_k$  as well as the  $\sigma_k$  (recall that  $\mathbf{M}$  is asymptotically the identity matrix).

A second alternative can be found in Kott (1990b). It replaces  $v$  with

$$v_B = v E_e[(t_R - T)^2 \mid \sigma^2 = \mathbf{s}^2] / E_e(v \mid \sigma^2 = \mathbf{s}^2) \quad (12)$$

Like  $v$ ,  $v_B$  is an asymptotically unbiased estimator for the randomization mean squared error to  $t_R$  under our assumptions. It is exactly model unbiased under the working model.

As we will see in the next section,  $v_B$  generalizes most easily among the bias-adjusted alternatives. Unfortunately, its implementation will often be messy in practice (especially when  $N < O(n^2)$ )

## 6. Other Sampling Designs

In practice, of course, there are many other sampling designs than the Poisson. We will focus first on other single-stage element-sampling designs and then move on to multi-stage designs.



## 6.1 Other Element-sampling Designs

It is not hard to show that  $t$  and  $t_R$  remain randomization consistent under the assumptions in equation (3) when  $\pi_{ik} \leq \pi_i \pi_k$  for  $i \neq k$ . This property is shared by most element-sampling designs with one notable exception – systematic sampling. We will restrict attention to designs where  $\pi_{ik} \leq \pi_i \pi_k$  when  $i \neq k$  for the remainder of the subsection.

The desirable model-based properties of  $t_R$  and  $v$  likewise are unchanged when we move from Poisson sampling to an alternative design with  $\pi_{ik} \leq \pi_i \pi_k$ . The model unbiasedness of  $t_R$  does not depend of the design at all. We do invoke a randomization-based property when asserting that the relative model bias of  $v$  is  $O_p(1/n^*)$  for a relatively large population. That property is unchanged in the expanded context under examination here.

Unfortunately,  $v$  is no longer necessarily asymptotically randomization-unbiased as an estimator for the randomization mean squared error of  $t_R$ . In many situations, it makes sense to replace  $v$  with

$$v^* = v + \sum_{i,k \in S (i \neq k)} [(\pi_{ik} - \pi_i \pi_k) / \pi_{ik}] (r_i / \pi_i) (r_k / \pi_k). \quad (13)$$

This variance/mean-squared-error estimator,  $v^*$ , can be shown to retain the model and randomization-based properties of  $v$  under Poisson sampling when the summation in equation (13) has only  $O(n^*)$  “cross” terms. This restriction assures that substituting for the model error term,  $e_k$ , and its randomization analogue,  $e_k$ , by  $r_k$  repeatedly in the

summation does not add appreciable bias (for example, the summation is at most  $O(1)$  under the model, while  $v$  itself is  $O(N^2/n^*)$ ). Observe that there are only  $O(n^*)$  cross terms under those stratified designs where the number of strata is large (and grows larger asymptotically) and the number of sample selections per stratum is small (and is bounded asymptotically).

Nothing is lost or gained in terms of the asymptotic properties of  $v^*$  by replacing some of all of the  $\pi_i$  in equation (13) by  $1/a_i$ . One potential gain is convenience as we shall see.

Simple random sampling is a special case where the summation in equation (13) has  $O([n^*]^2)$  terms, but  $v^*$  retains its desirable properties. That is because it also has this form:

$$\begin{aligned} v^* &= v + \sum_{i,k \in S (i \neq k)} [(\pi_{ik} - \pi_i \pi_k) / \pi_{ik}] (r_i / \pi_i) (r_k / \pi_k) \\ &= v + [(1 - n/N) / (n - 1)] (N/n)^2 \left[ \left( \sum_{i \in S} r_i \right)^2 - \sum_{i \in S} r_i^2 \right] \end{aligned} \quad (14)$$

When there is a vector  $\mathbf{q}$  such that  $c_k \mathbf{x}_k \mathbf{q} = 1$  for all  $k \in S$ , then  $\sum_S r_i = 0$ , and  $v^*$  simplifies further. In fact, we can drop the  $\sum_S r_i^2$  term entirely because it is asymptotically ignorable ( $v$  is  $O_p(N^2/n)$ , while  $(N/n)^2 \sum_S r_i^2 / (n - 1)$  is  $(N^2/n) O_p(1/n)$ ). Alternatively, we can partially acknowledge its presence in an *ad hoc* fashion by setting  $v^* = nv / (n - 1)$ .

Small-sample adjustments to  $v^*$  in equation (13) are possible using the approach in equation (12). For simple random sampling, when the  $\mathbf{x}_k$  are such that  $\sum_S r_i = 0$ , the

method in equation (11) is also viable as is the fully model-unbiased method that computes  $r_{(2)}$ .

It is a simple matter to generalize equation (14) to cover stratified simple random sampling. That generalization does not always have a simple form. When there are few strata and a large number of selections per each stratum, it makes sense to expand the x-vector so that  $\sum r_i = 0$  where the summation is over the sample *within each stratum*. When this equality holds for each stratum,  $v$  itself will serve as a good variance estimator in both the model and randomization-based senses. It can be small-sample-bias adjusted using the method in either equation (11) or (12) or the fully model-unbiased method.

When there are many strata and few selections per stratum, then equation (14) can be replaced by an alternative

$$\begin{aligned}
 v' &= v + \sum_{h=1}^H \sum_{i,k \in S_h (i \neq k)} [(\pi_{ik} - \pi_i \pi_k) / \pi_{ik}] (a_i r_i a_k r_k) \\
 &= v + \sum \sum (1 - n_h / N_h) / (n_h - 1) (a_i r_i a_k r_k) \\
 &\approx \sum_{h=1}^H \left\{ \sum_{i \in S_h} (a_i^2 - a_i) r_i^2 + \sum_{i,k \in S_h (i \neq k)} [1 - 1 / (\downarrow a_i \downarrow a_k)] / (n_h - 1) (a_i r_i a_k r_k) \right\} \\
 &= \sum_{h=1}^H (n_h / [n_h - 1]) \left\{ \sum_{i \in S_h} (a_i^2 - a_i) r_i^2 - [(\sum_{i \in S_h} a_i r_i)^2 - (\sum_{i \in S_h} \downarrow a_i r_i)^2] / n_h \right\}, \tag{15}
 \end{aligned}$$

where  $S_h$  is the sample within stratum  $h$ , and  $n_h / N_h$  the sampling fraction (selection probability) is that stratum.

Equation (15) remains more homely than we might have liked. When finite population correction can be ignored (when all  $N_h \gg n_h$ ), however, the equation simplifies nicely:

$$v^{**} = \sum (n_h/[n_h - 1]) \{ \sum (a_i r_i)^2 - ( \sum a_i r_i)^2 / n \}. \quad (16)$$

This same variance equation we can use in practice for many stratified designs with *unequal* selection probabilities within each strata when all  $\pi_k \ll 1$ . The key is that  $\pi_{ik}/\pi_i\pi_k$  needs to be  $(n_h - 1)/n_h + O(n^*/N)$  for all unequal  $i$  and  $k$  in stratum  $h$ .

Again, small-sample-bias adjustment can be applied using the method in either equation (11) or (12). The fully model-unbiased method, however, is no longer viable.

It is of some interest to note that  $v^{**}$  in equation (16), since it ignores finite population correction, expresses the model variance of  $t_R$  as estimator for  $\sum_U \mathbf{x}_k \beta$ .

## 6.2 Multi-stage Sampling Designs

In a multistage design, a cluster of elements called a primary sampling unit (PSU) is first selected *without* replacement, then probability samples of elements are selected independently within each PSU. We will not formally address with-replacement sampling, either real or fictitious, here.

Let  $n_i$  denote the number of PSU's in the sample, and  $n_j$  the number of elements subsampled in each PSU. If  $n_j$  is bounded for all  $j$  as  $n^*$  grows arbitrarily large (so that

$O(n_l) = O(n^*)$  , it is a simple matter to show that all  $t_R$  remains randomization consistent as long as  $\pi_{ljg} < \pi_{lj}\pi_{lg}$  when  $j \neq g$ , where  $\pi_{lj}$  is the selection probability of PSU  $j$  and  $\pi_{ljg}$  is the joint selection probability of PSU's  $j$  and  $g$ .

It is common to estimate the variance of  $t_R$  with the multistage analogue of equation (16):

$$v^{**} = \sum_{h=1}^H (n_{lh}/[n_{lh} - 1]) \left\{ \sum_{j \in S_{1h}} \left( \sum_{i \in S_{hj}} a_i r_i \right)^2 - \left( \sum_{j \in S_{1h}} \sum_{i \in S_{hj}} a_i r_i \right)^2 \right\}, \quad (17)$$

where  $h$  denotes a stratum of PSU's,  $n_{lh}$  the number of sampled PSU's in stratum  $h$ ,  $S_{1h}$  the set of sampled PSU's in  $h$ , and  $S_{hj}$  the set of subsampled elements from PSU  $hj$ .

The estimator in equation (17) is asymptotically randomization unbiased for the randomization mean squared error of  $t_R$  when  $\pi_{ljg}/(\pi_{lj}\pi_{lg}) - (n_{lh} - 1)/n_{lh}$  is ignorable small. It is easy to see that it is also asymptotically model unbiased for the model variance of  $t_R$  as an estimator for  $\sum_U \mathbf{x}_k \beta$ . In fact, we can generalize the error structure of the model somewhat. Instead of requiring  $E(e_i e_k)$  to be zero when  $i \neq k$ , we now require only that this correlation be bounded when  $i$  and  $k$  are from the sample PSU. When  $i$  and  $k$  are from different PSU's,  $E(e_i e_k)$  is again assumed to be zero. Observe that under this structure both  $E_e(v^{**})$  and  $E_e[(t_R - \sum_U \mathbf{x}_k \beta)^2]$  are (asymptotically in the case of the former) equal to  $\sum E_e[(\sum_{i \in S(hj)} a_i e_i)^2]$ , where the first summation is over all the PSU's in the first-stage sample (and  $S(hj) = S_{hj}$ ).

The method in equation (12) can again be applied in an attempt to remove the small-sample model bias of  $v^{**}$ . This assumes, however, that all the  $E(e_i e_k) = 0$ .

Alternatively, we can replace  $\sigma^2 = \mathbf{s}^2$  with a more complex assumption about the error structure of the  $\epsilon_k$ . As before, if this working model is correct up to a scalar, the variance estimator is model unbiased. Otherwise its relative bias is  $O(1/n^*)$ .

### 6.3 Effective Degrees of Freedom

The real world is not asymptotic. The expected element sample size may be large in practice, but with multistage sampling,  $n_i$  will be less so. Kott (1994) discusses how to calculate the relative model variance of  $v^{**}$  under ideal conditions; that is to say, when we the  $\epsilon_k$  are normally distributed with  $E(\epsilon_i \epsilon_k) \propto s_{ik}$  for a hypothesized set of  $s_{ik}$  (note:  $s_{ik} = 0$  when  $i$  and  $k$  are from different PSU's).

If  $v^{**}$  had a chi-squared distribution, its degrees of freedom would be related to its relative variance:  $F = 2/\text{relvar}(v^{**})$ . With this in mind, Kott proposes the following Satterthwaite-like calculation for the effective degrees of freedom of  $v^{**}$ :

$$F = \frac{\left( \sum_{h=1}^H \sum_{j \in S_{1h}} v_{hj} \right)^2}{\sum_{h=1}^H \left\{ \sum_{j \in S_{1h}} v_{hj}^2 + \sum_{g \neq j \in S_{1h}} v_{hj} v_{hg} / (n_h - 1) \right\}}, \quad (18)$$

where  $v_{hj} = E_e \left[ \left( \sum_{i \in S(hj)} a_i \epsilon_i \right)^2 \middle| E(\epsilon_i \epsilon_k) = s_{ik} \right]$ . The idea is that one should construct a confidence interval for  $t_R$  assuming the pivotal,  $t_R / \sqrt{v^{**}}$ , has Student's  $t$  distribution with  $F$  degrees of freedom. Kott shows that attempting to estimate  $F$  from the sample without assumptions about the structure of the  $\epsilon_k$  is not advisable.

The same computation of  $F$  could be applied if  $v^{**}$  were bias-adjusted using the method in equation (12). Bell and McCaffrey (2001) have pointed out that the determination of  $F$  above ignores the distinction between  $e_k$  and  $r_k$ . They offer a theoretically superior alternative, which, unfortunately requires  $v^{**}$  to have an unstratified form:

$$v^{**} = \sum_{h=1}^H \sum_{j \in S_{1h}} \left( \sum_{i \in S_{hj}} a_i r_i \right)^2.$$

They also propose different methods of removing the model bias of  $v^{**}$  from those discussed here. These methods likewise require a working model about the variance/covariance matrix for the  $e_k$ .

## 7. Concluding Remarks

We have attempted here to commit to paper an idea this author has long espoused in public; namely, that the dominant model-assisted (randomization-based) survey-sampling paradigm, although fruitful in many ways, should be supplanted by a randomization-assisted model-based one. That is because inference should be based on the sample actually observed rather than averaged over all potential samples. Randomization-based methods provide some protection against inevitable model failure, but that protection relies on invoking the powers asymptotic properties in a finite world.

Many have been unwilling to use asymptotic statistics at all in the service of

survey sampling. This is because the principal goal of survey sampling is the estimation of *finite* population parameters. The real issue, however, is not whether the population can be viewed as large, but whether the sample can. It is precisely because samples are often very large in survey sampling, that the apparently exclusive use of randomization methods dominated its practice for so long.

Notice that the modifier “apparently” in the last sentence. Model were always there, lurking in the background of strictly randomization-based survey sampling, helping practitioners choose among estimation strategies. The model-assisted paradigm took models out of the shadows and awarded them a formal place in survey theory and practice. Still, real inference was deemed to be related to the randomization distribution of an estimator, not its model distribution.

In the model-assisted paradigm, one chooses among randomization-consistent estimation strategies by hypothesizing a reasonable and practical model, restricting attention to model-unbiased estimators, and selecting that strategy with the minimum model-expected randomization mean squared error.

We have argued that this selection routine makes even more sense under a randomization-expected model-based paradigm. Moreover, the weight-residual variance estimator from Särndal, Swensson, and Wretman (1989) actually does a better job estimating model variance than randomization mean squared error, yet that is the variance/mean squared error estimator given in Särndal et al.’s 1992 text book on model-assisted methods.

By paying closer attention to the asymptotics, we improved a bit on their variance estimator in the text, concentrating first on those situations where finite population



correction matter and then on cases where the sample itself is not very large. As the population and then the sample become less large, it was necessary to make more assumptions about the error structure of the model. This may be regrettable, but relying instead on randomization-based properties, which are asymptotic in nature, makes little sense. Indeed, as we have seen, models can help us ferret out just when the sample may be too small to assert the asymptotic normality of the pivotal  $(t_R / \sqrt{v^*} \text{ or } t_R / \sqrt{v^{**}} \text{ when finite population correction is ignorable})$ , a common and often unjustified practice.

There are many topics we have not had the space to address here. Kott and Bailey (2000) discuss a method for drawing a multipurpose sample under the randomization-assisted model-based paradigm. Kott (1998) shows that the jackknife and balance-repeated-replication variance estimators share the asymptotic model-based properties of the weighted-residual variance estimator when finite population correction can be ignored.

We have not discussed techniques for handling the impact of nonresponse and measurement error, two areas where the use of model-based methods is already widely accepted. Another area where model-based methods are widely used is in small-domain estimation. This is because sample sizes can be too small for randomization-based theory to have much virtue. There, the incorporation of randomization-based principles is problematic. On the one hand, for the small domains sampling weights are, at best, a nuisance. On the other, when small domains are aggregated together, the result is based on a large enough sample than randomization-based principles can offer some protection against model failure. Although not described that way, the

approach to small-domain estimation in You and Rao (2001) is consistent with the randomization-assisted model-based paradigm.

Opsomer and Breidt (2000) have developed a promising randomization-consistent local-polynomial-regression estimator. Their variance estimator, however, does not have the desirable asymptotic model-based property discussed here. That failing should be corrected. Finally, model-assisted papers on strategies using multi-phase sampling are appearing in the literature with increasing frequency (see, for example, Hidiroglou and Särndal, 1998) . Just how to handle such designs from a randomization-assisted viewpoint needs to be addressed.

## References

- Bell, R. and McCaffrey, D. (2001). Bias reduction in standard errors for linear regression with multi-stage samples. ASA Proceedings of the Section on Survey Research Methods, forthcoming.
- Breidt, F.J. and Opsomer, J. D. (2000). Local polynomial regression in survey sampling. *Annals of Statistics*, 28, 1026-1053.
- Brewer, K.R.W. (1963). Ratio estimation and finite populations: some results deductible from the assumption of an underlying stochastic process. *Australian Journal of Statistics*, 5 93-105.
- Brewer, K.R.W. (1994). Survey sampling inference: some past perspectives and present prospects. *Pakistan Journal of Statistics*, 10(1)A 213-233.
- Chew, V. (1970). Covariance matrix estimation in linear models. *Journal of the American Statistical Association*, 65 173-181.
- Deville, J-C. and Särndal, C-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87 376-382.
- Estevao, V.M., and Särndal, C-E. (2000). A functional form approach to calibration. *Journal of Official Statistics*, 16, 379-399.
- Hidiroglou, M.A, and Särndal, C-E. (1998). Use of auxiliary information for two-phase sampling. *Survey Methodology*, 24, 11-20.
- Isaki, and Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

- Kott, P.S. (1990a). Estimating the conditional variance of a design consistent regression estimator. *Journal of Statistical Planning and Inference*, 24, 287-296.
- Kott, P.S. (1990b). Recently proposed variance estimators for the simple regression estimator. *Journal of Official Statistics*, 6, 451-454.
- Kott, P.S. (1998). A model-based evaluation of several well-known variance estimators for the combined ratio estimator. *Statistica Sinica*, 8, 1165-1173.
- Kott, P.S. (2002). A practical use for instrument-variable calibration. *Proceeding of the Section on Survey Research Methods*, forthcoming.
- Kott, P.S. and Bailey, J. T. (2000). The theory and practice of maximal Brewer selection. *Proceedings of the Second International Conference on Establishment Surveys*, Invited papers, 269-278.
- Kott, P.S. and Brewer, K.R.W. (2001). Estimating the model variance of a randomization-consistent regression estimator. *ASA Proceedings of the Section on Survey Research Methods*, forthcoming.
- Särndal, C-E, Swensson, B., and Wretman, J. (1989). The weighted residual technique for estimating the variance of the general regression estimator of a finite population total. *Biometrika*, 76, 527-537,
- Särndal, C-E, Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer.
- Singh, A.C. and Mohl, C.A. (1996). Understanding calibration estimators in survey sampling. *Survey. Methodology*, 22 2, 107-115.
- Wu, C.F.J. and Deng, L.Y. (1983). Estimation of variance of the ratio estimator: an empirical study. In *Scientific Inference, Data Analysis and Robustness*, ( Eds.

G.E.P. Box, *et al.*) New York: Academic Press, 245-277.

You, Y. and Rao, J.N.K. (2001). Pseudo EBLUP estimation of small area means using survey weights.